Semantische Textanalytik I: Grundlagen des Information Retrieval

Seminar im Modul M-GSW-09
WiSe 2016/17

Prof. Dr. Udo Hahn

Lehrstuhl für Angewandte Germanistische Sprachwissenschaft / Computerlinguistik

Institut für Germanistische Sprachwissenschaft

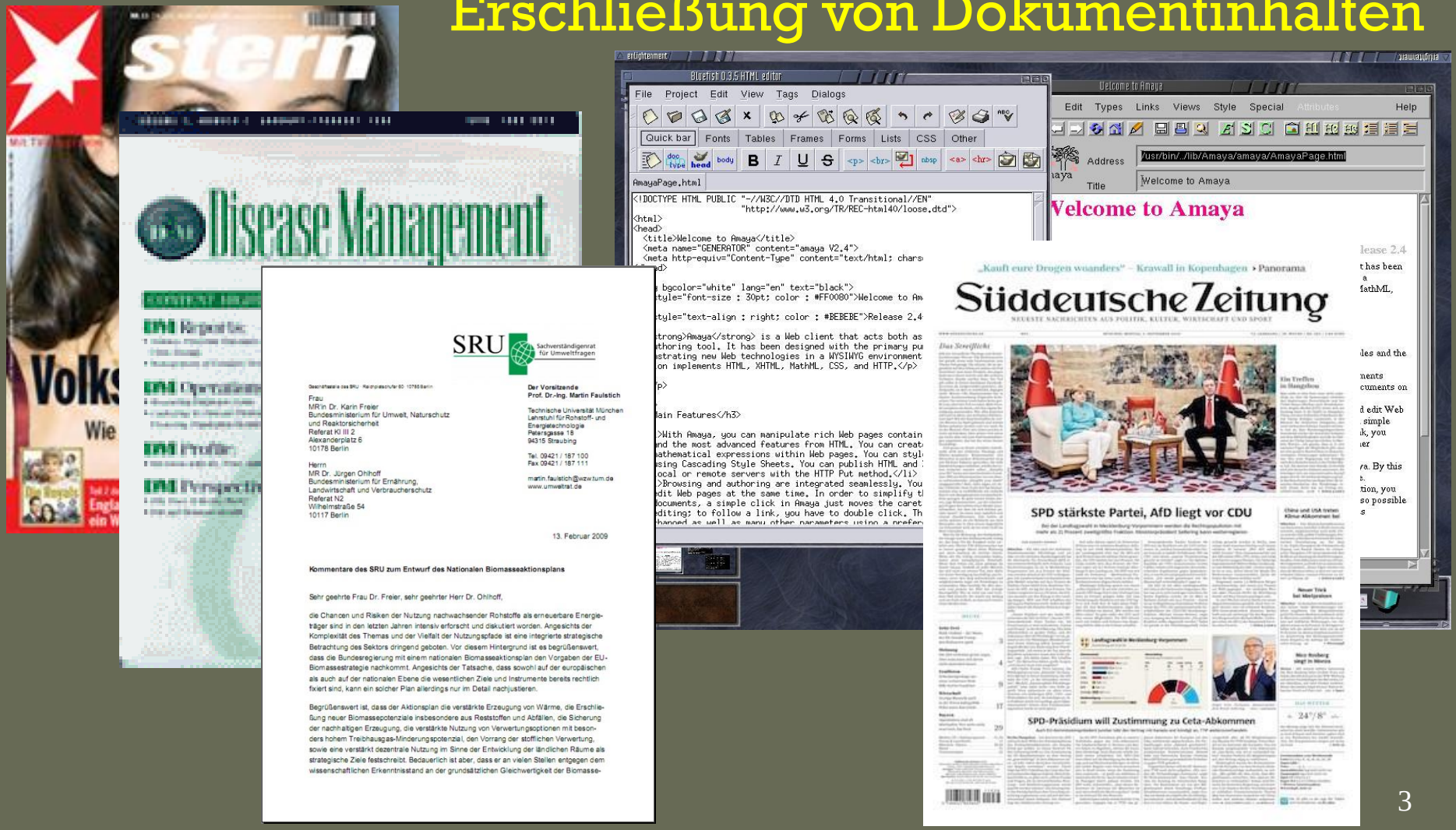
Friedrich-Schiller-Universität Jena

http://www.julielab.de

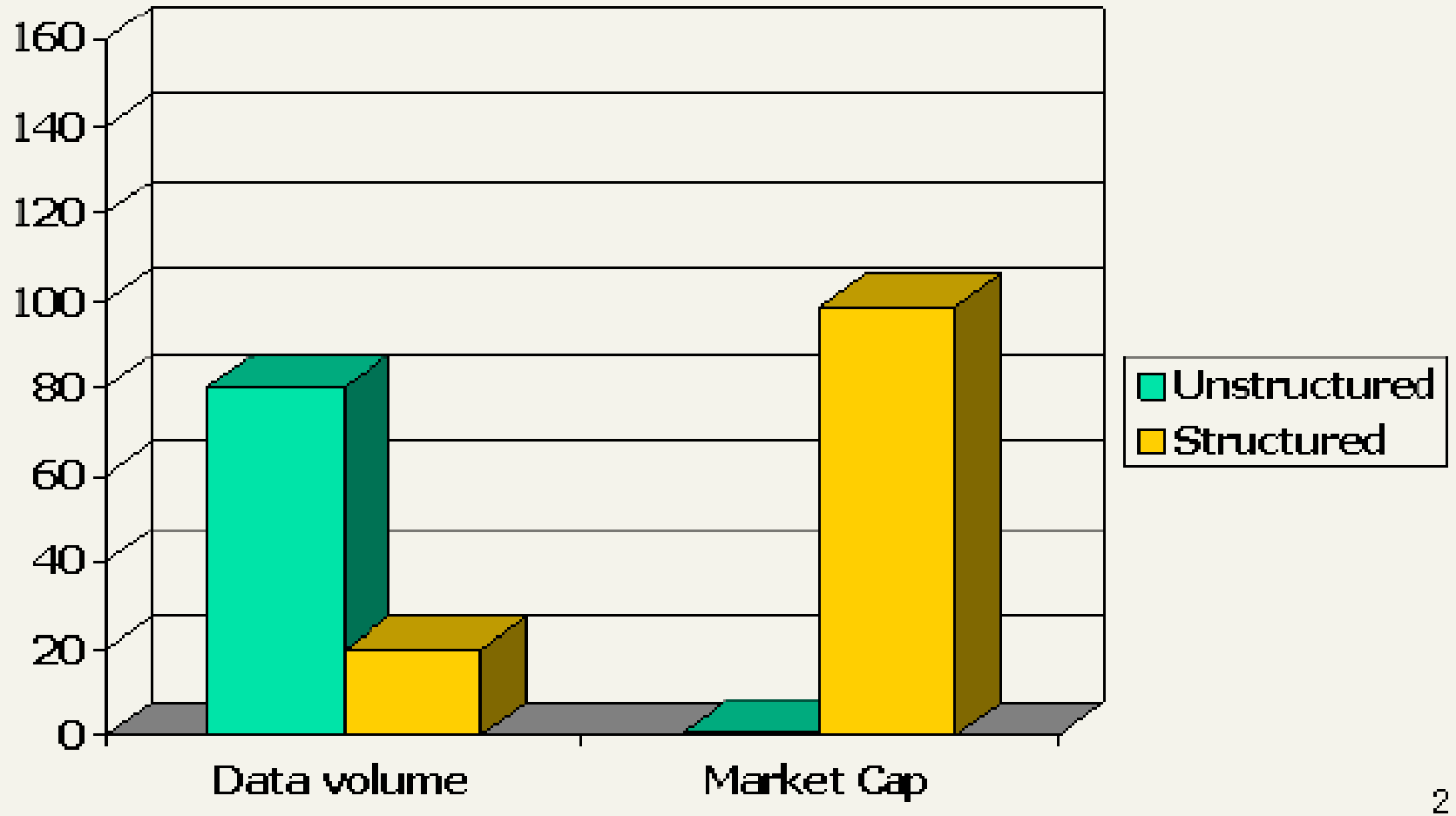
Allgemeine Hinweise

- Termin: Do, 16-18h (Humboldt 11, SR 115)
- Materialien im Netz
  - http://www.julielab.de ☞„Students“
- Sprechstunde: Mi, 12-13h (FüG 30, R 203)
- Email: udo.hahn@uni-jena.de
- Fachliteratur: durchgängig in Englisch

Sammeln von Dokumentkollektionen vs. Erschließung von Dokumentinhalten



Strukturierte vs. unstrukturierte Daten (1996)



Quelle: Prabhakar & Raghavan, Verity (2002)

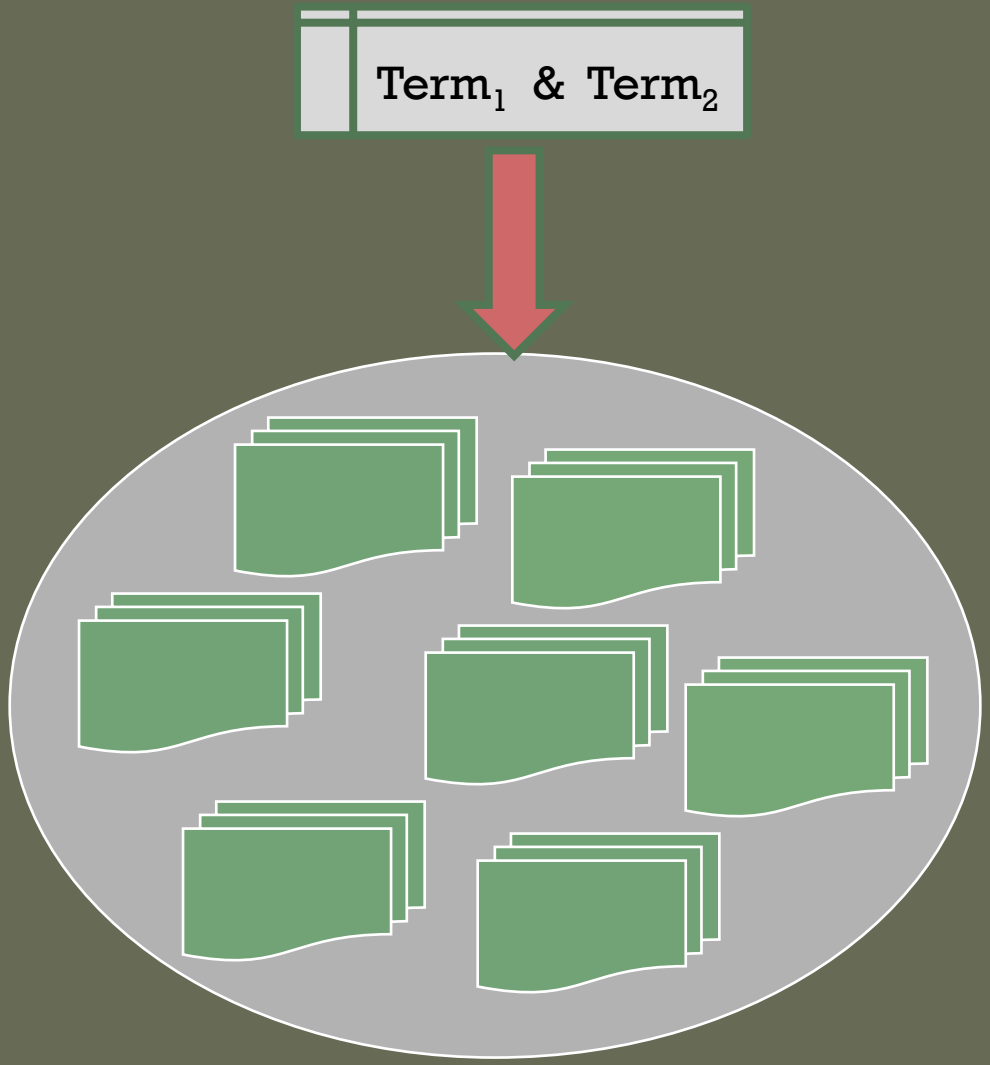
Strukturierte vs. unstrukturierte Daten (2006)

Google

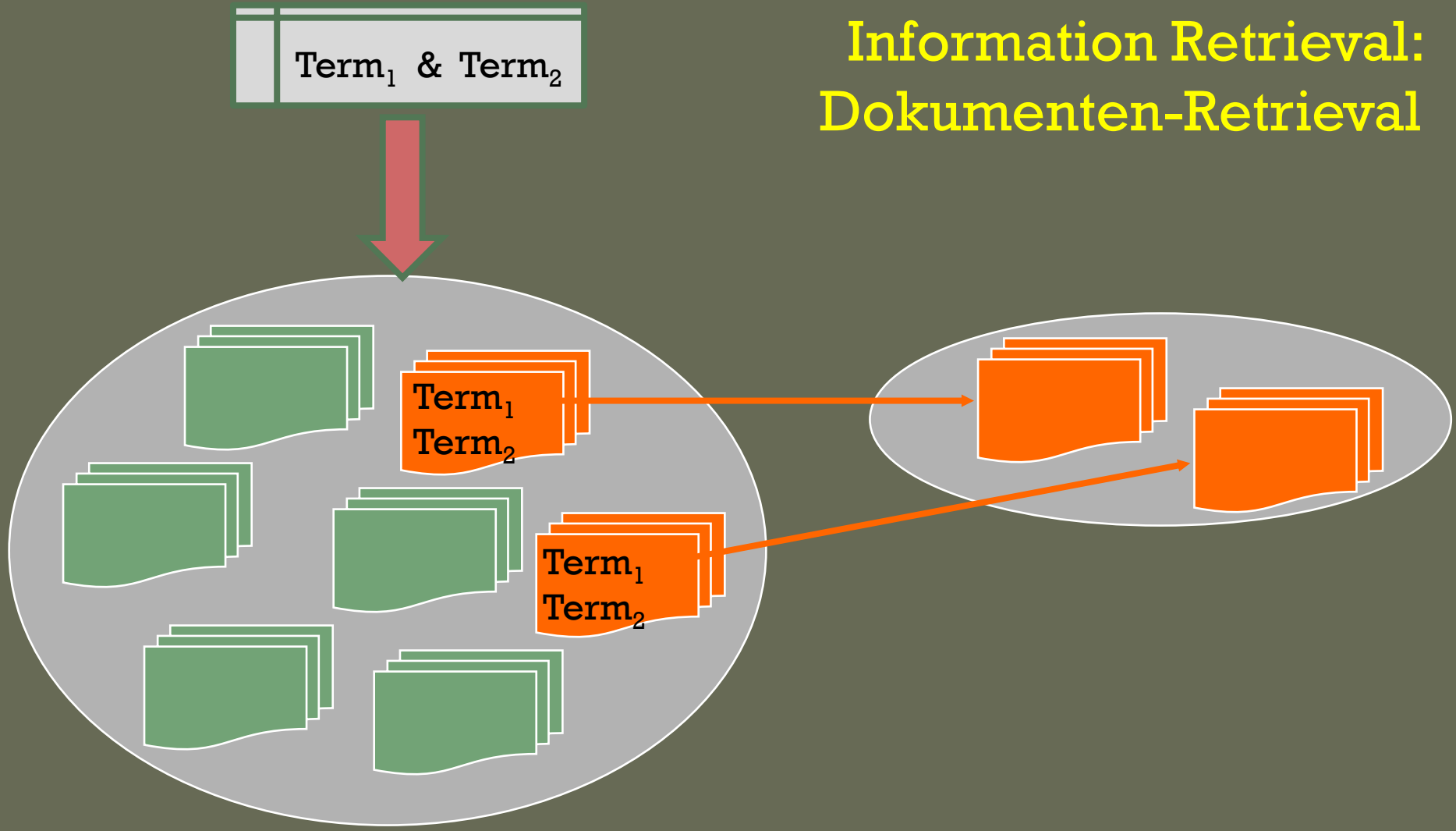
YAHOO!

Ask

Information Retrieval: Dokumenten-Retrieval



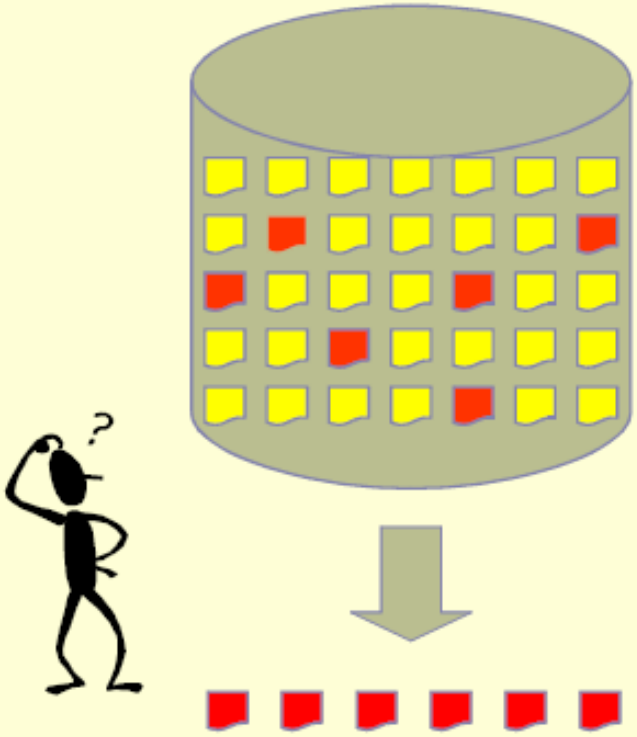
Information Retrieval: Dokumenten-Retrieval



Ad hoc Retrieval vs. Routing

Ad-hoc retrieval

One time queries (e.g. Web search)



Filtering/Routing

Constant search profile (e.g. Spam filtering)

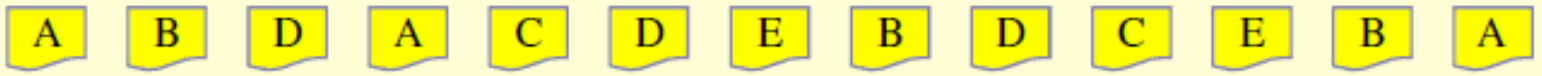
Textkategorisierung: Clustering vs. Themenerkennung

- Categorization/Clustering:

  Group documents into predefined classes/ adaptive clusters

- Topic Detection and Tracking:

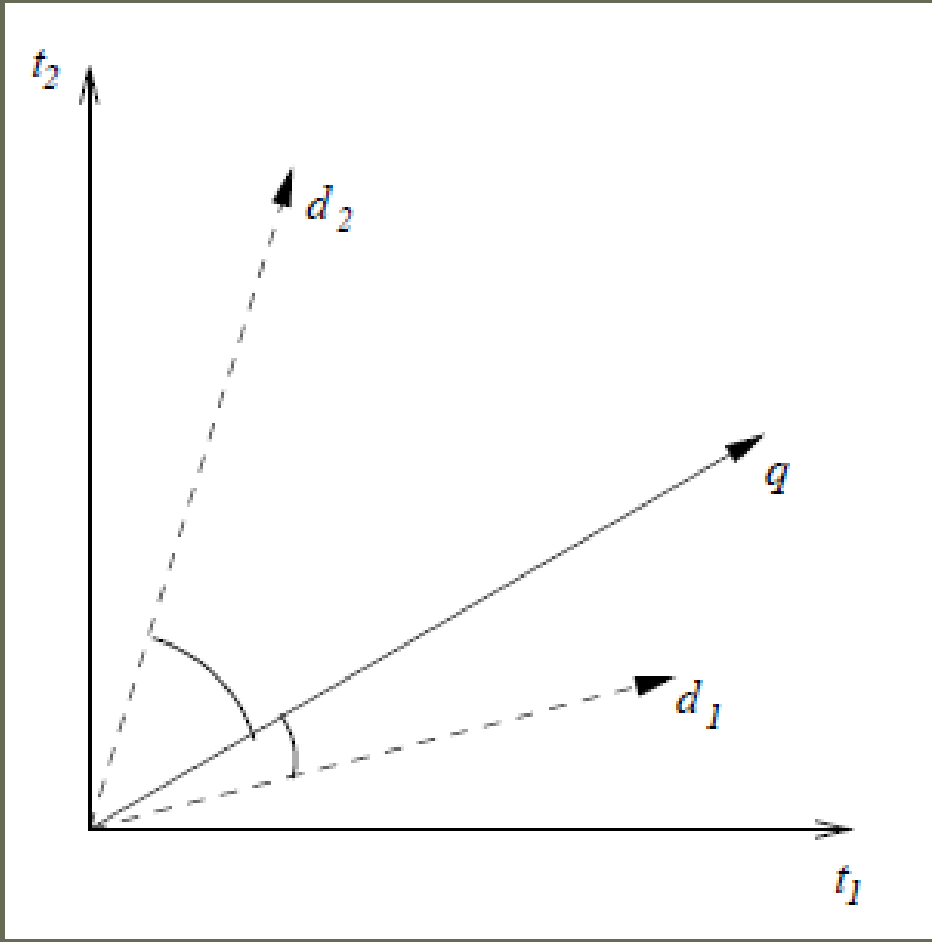
  Cluster news in stream



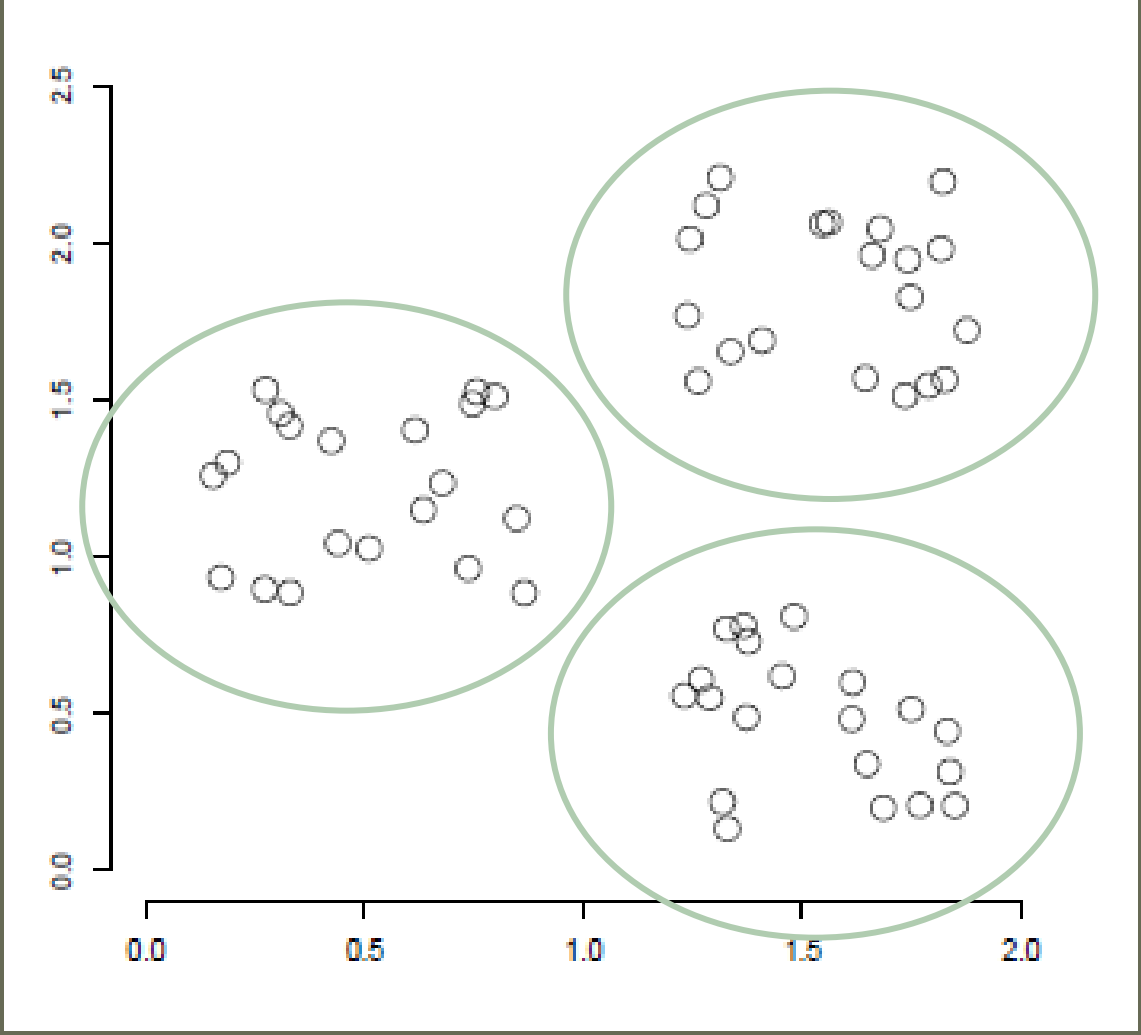
Information Retrieval Modelle

- Klassisches Boole‘sches Retrieval-Modell
  - Boole‘sche Operatoren (UND, ODER, NICHT)
  - Textstruktur-Operatoren (NEXT, IN-UNIT)
- Vektorraum-Modell
  - Dokument-Term-Matrizen
  - Ähnlichkeitsmaße (Cosinus, Jaccard usw.)
  - Clustering-Verfahren
- Probabilistische Modelle
  - Annahme spezieller Verteilungen von Termen

Vektorraum-Modell



Clustering



Experimentelle Rahmenbedingungen

- „Relevanz“
- Retrieval-Metriken
  - Recall, Precision, F-Score
- Retrieval-Wettbewerbe
  - TREC (seit 1992, jährlich)
    - Dutzende von Tracks: Ad hoc, Routing, QA, video, …
  - CLEF (seit 2000, jährlich)
    - Multi/cross-lingual, nicht-englische Sprachen

Seminarleistungen

- Vortrag (mündlich)
  - 1-stündig
  - Elektronische Version (PDF, PPT) verfügbar machen
- Referat (schriftlich)
  - 15-20 Seiten Kerntext (mit Standardformaten)
  - Elektronische Version (PDF, DOC) verfügbar machen
  - Eidesstattliche Erklärung zur Eigenautorenschaft
    - Wir prüfen mit Plagiatserkennungs-Software
  - Abgabe: Anfang März 2017

Bemerkungen zu Referaten

- Aufbaumuster:
  - Deck- bzw. Titelblatt mit vollständigen Angaben
  - Inhaltsverzeichnis
  - Einführung ins Thema, Motivation
  - Themenabhandlung: grundlegende Formalisierungen, Verfahrensbeschreibungen (Algorithmen), Systemfunktionalitäten, Ressourcenmerkmale, Experimente/Evaluationen usw.
  - Fazit mit kritischer Würdigung, offene Probleme ansprechen
  - Bibliographie
- Zitationen:
  - Alle verwendeten Quellen zitieren
    - Mit einem bibliographisch korrektem Zitat die jeweilige Quelle eindeutig beschreiben
    - Fachartikel nicht mit **http**://...**foo**.**pdf**-Link zitieren
    - Online-Quellen mit URLs und Datum des letztem Zugriffs
  - **Wikipedia** ist keine zitierfähige wissenschaftliche Quelle !
- Eigenleistungen (Literatur, Beschäftigung mit konkreten Ressourcen/Systemen usw.) sind sehr erwünscht→unabdingbar !

Wege zum Vortrag und Referat

- Email: Anmeldung von drei nach fallender Priorität geordneten Themenwünschen
  - First-come, first-served
- Email: Themenvergabe durch Dozenten
- Erste Literaturhinweise als „Saat“ nach Bestätigung der Themenauswahl
- Themenbearbeitung durch Referenten
  - Mündlicher Vortrag zum vereinbarten Termin
  - Schriftliches Referat (unter Einhaltung der organisatorischen Verabredungen) zum vereinbarten Termin

Grundlegende Literatur

- * Baeza-Yates, Ricardo A., & Berthier Ribeiro-Neto(1999). *Modern Information Retrieval*. New York, N.Y.: ACM Press/Addison-Wesley.
- * Manning, Christopher D., Prabhakar Raghavan, & Schütze, Hinrich (2008). *Introduction to Information Retrieval*. New York/NY, USA: Cambridge University Press
- Büttcher, Stefan, Charles L. A. Clarke, & Gordon V. Cormack (2010). *Information Retrieval: Implementing and Evaluating Search Engines*. Cambridge/MA: MIT Press
- Croft, W. Bruce, D. P. Metzler, & T. Strohman (2009). *Search Engines: Information Retrieval in Practice*. Reading/MA etc.: Addison-Wesley Publishing Company

Wichtige Zeitschriften

- Information Processing & Management
- Journal of the American Society for Information Science
- Journal of Documentation
- Foundations and Trends® in Information Retrieval
- Information Retrieval Journal
- Annual Review of Information Science and Technology (ARIST)

Wichtige Konferenzen

- SIGIR – Proceedings of the nth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval
- CIKM – Proceedings of the nth ACM Conference on Information and Knowledge Management

Ablaufplan

20.10.	Hahn
27.10.	Hahn – Themenvergabe
03.11.	---
10.11.	---
17.11.	---
24.11.	---
01.12.	---
08.12.	---
15.12.	---
22.12.	---
05.01.	Reiter: Grundlagen IR: Boolesch, VR
12.01.	Glaß: Web-Retrieval: PageRank u.a.
19.01.	Ziemer: QA
26.01.	
02.02.	---